

VITA in Use: Technology Watch

Zachary Jacobson, S.G. McIntyre, Tiit Romet

Applied Research and Analysis Directorate
400 Cooper Street, Postal Locator: 4602B
Ottawa, Ontario K1A 0K9
CANADA

INTRODUCTION

The potential wealth to be discovered from mining large data sets or electronic text files for nuggets of knowledge is an alluring prospect. Textual displays, however, do not lend themselves easily to the task as the results still require reading and analysis before the analyst/user can acquire new knowledge from the information therein. Visual displays, on the other hand, attempt to perform some of the preliminary cognitive analysis.

VITA, a “Visual Interface for Text Analysis”, is a 3-dimensional paradigm to identify the relations found among meanings or concepts represented in the elements in large text corpora. The paradigm has been realized as a working software application used to direct computer-based document searches. It allows a user, via mouse and keyboard action, to interact with search mechanisms – e.g. search engines on the Internet, such as Google and AltaVista - to present visually the sets and relationships of documents. VITA has control features that allow visual clustering of like documents, thus enabling quick refinement of the search process. The visual features of VITA also support the observation and investigation of the, sometimes unexpected, relationships among documents.

VITA can also be used to help in reducing document search complexity. Originally conceptualized as a response to the problem of comprehending the results of large computer-based document searches, VITA has the potential for broader applications in text mining and knowledge discovery. One such application is Technology Watch.

“Tech Watch” is a methodology used to identify technology trends and make strategic investments in science and technology research and procurement. It looks for strengths, gaps and trends within the national and international technology scenes to incorporate into long-term planning. Defence R&D Canada is investigating various tools for potential use in their Tech Watch project and proposed that VITA demonstrate its applicability for this problem. This paper shows how VITA responded in the practical situation and discusses how it might be adapted to future Tech Watch problems.

FEATURES OF VITA

Built as a “bolt-in” application that accepts a standard search engine and generates an interactive display, VITA parallels the IST-05 Reference model closely [see figure 1].

Paper presented at the RTO IST Workshop on “Massive Military Data Fusion and Visualisation: Users Talk with Developers”, held in Halden, Norway, 10-13 September 2002, and published in RTO-MP-105.

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE 00 APR 2004		2. REPORT TYPE N/A		3. DATES COVERED -	
4. TITLE AND SUBTITLE VITA in Use: Technology Watch				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Applied Research and Analysis Directorate 400 Cooper Street, Postal Locator: 4602B Ottawa, Ontario K1A 0K9 CANADA				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release, distribution unlimited					
13. SUPPLEMENTARY NOTES See also ADM001665, RTO-MP-105 Massive Military Data Fusion and Visualization: Users Talk with Developers., The original document contains color images.					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 28	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

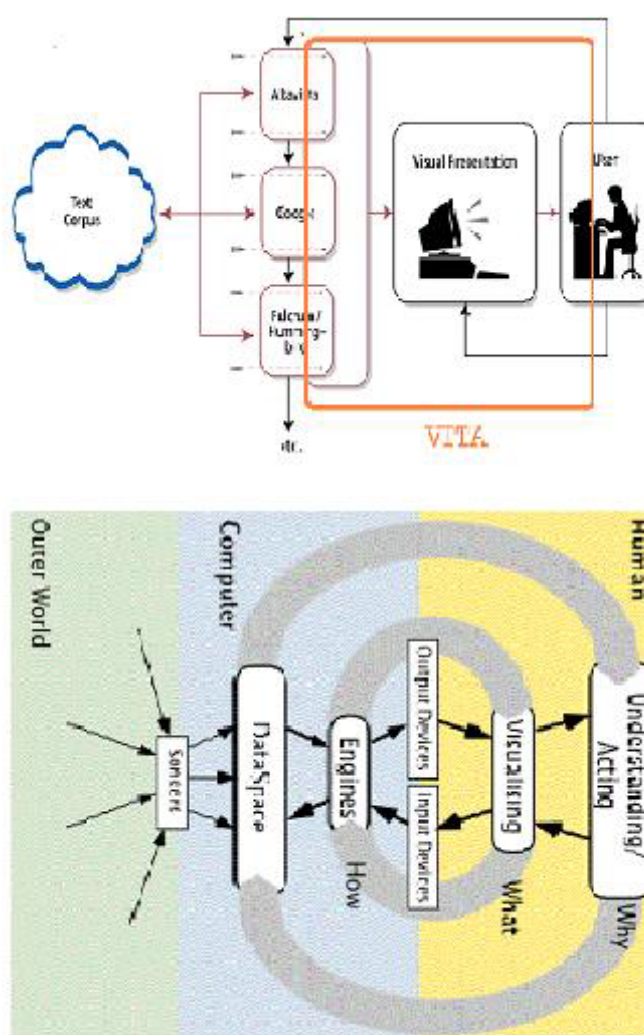


Figure 1: VITA Follows The Standard IST-05 Reference Model.

There are four suggested functions, as per the IST-05 Reference Model, that a visualization system such as VITA could be applied to, controlling or monitoring, searching, exploring (screening) and alerting. The VITA system has shown varying utility based on the function or task that is being performed, with the initial results reflecting the greatest use in the exploring or screening of a data set. Each function is also influenced by other factors such as the specificity and applicability of the query set; the detail and nature of the tagging of the elements of a dataset; and human factors related to subject matter expertise, an understanding of the task, and an understanding of the strengths and limitations of the visualization tool.

VITA allows a user, via mouse and keyboard action, to interact with search mechanisms such as search engines on the Internet, e.g., Google or AltaVista, to present visual displays of documents sets of potential interest. VITA has control features that allow visual clustering of like documents, thus enabling quick refinement of the search process. The visual features of VITA also support the observation and investigation of, sometimes unexpected, relationships among documents.

VITA has been developed as a research testbed to identify better methods of visualizing relevant document clusters and identifying their relationships. As such, there have been various prototypes created as different ideas emerge. VITA presently exists in two versions: VITA-delta and VITA-epsilon. VITA-Delta is written in Visual Basic 6.0 and is a more elaborate, but scale-limited research-oriented

Epsilon is written in C++ and is faster and more robust. As each has its strengths, both were employed in the Tech Watch project.

TESTING VITA FOR TECH WATCH

Two approaches to evaluate VITA as a potential Tech Watch tool were determined. The first was to employ a defence-technologies taxonomy to determine how it would be mapped in an open source context. The Canadian Defence R&D Program taxonomy was mapped to the U.S. Defence Technology Area Plan (DTAP). The entire taxonomy was then searched on a sub-set of Canadian sites. For this test, the VITA – epsilon version was used.

The results of the mapping, shown in figure 2, display the relationships between the two taxonomies. The selected example, missiles, is a component of a number of areas of research. It is easy to see, in this figure, which technologies are well addressed, and perhaps more importantly, which are not well addressed or not addressed at all. In practice, VITA serendipitously shows the “holes” at as easily as the “doughnut”.

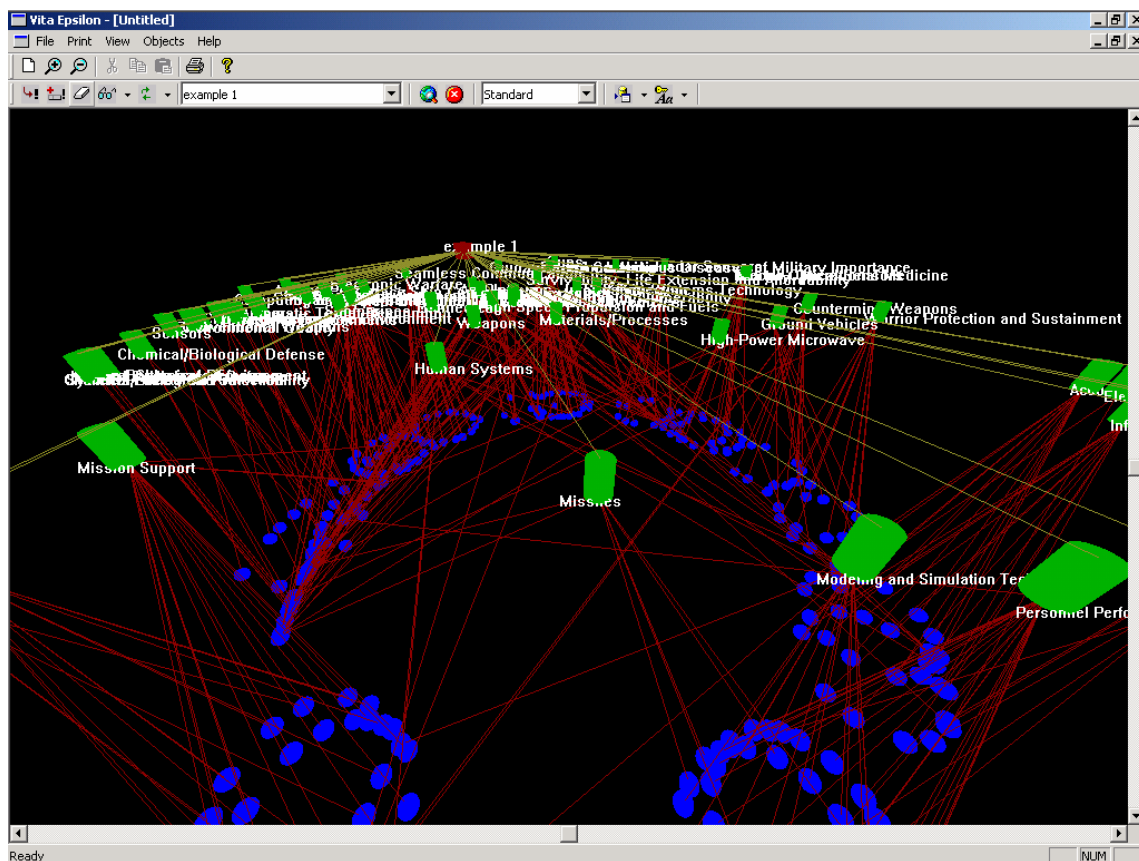


Figure 2: The more robust “production” epsilon version of VITA was used to map the defence taxonomy against all existing Canadian Defence R&D projects. Green cylinders show the technologies and concepts in the taxonomy. Blue spheres show specific R&D projects that relate to those concepts. Linkages are shown as red lines.

The second test used VITA-delta to search for a specific example taken from the taxonomy. Canada’s National Research Council website was queried for documents connecting fuel cells and air weapons systems.

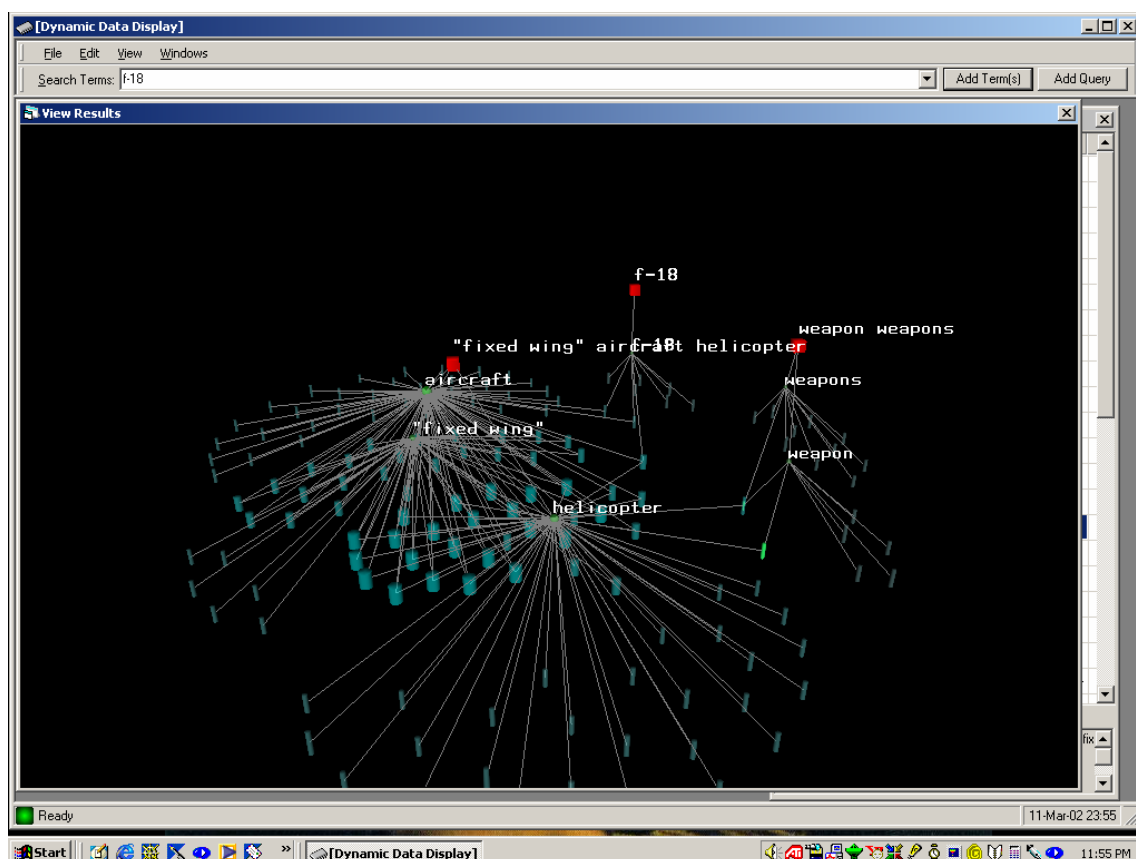


Figure 3: VITA session showing a series of progressively more refined queries concerning aircraft uses of fuel cells and related work and interest at the National Research Council of Canada.

The taxonomy shows:

Air Platforms (DTAP broad term)

- Air Weapons Systems (CA Thrust 13e)
- Aircraft/Weapon System Compatibility (CA Project 13ec)

Fixed-Wing Vehicles (DTAP narrow term)

Rotary-Wing Vehicles (DTAP NT)

Integrated High-Performance Turbine Engine Technologies (DTAP NT)

Aircraft Power (DTAP NT)

- Advanced Power Sources (CA Project 13gf)
- Advanced Portable Fuel Cells (CA Project 13gj)

High-Speed Propulsion and Fuels (DTAP NT)...

We derived search terms for queries and probes were derived from this segment. The search proceeded as follows:

The first query was

Aircraft, "fixed-wing", helicopter...

This yielded a filled field of returns. But the query

Weapon weapons system...

showed only a slight intersection with any part of the previous query, and none with *fixed wing*.

The two intersects with *helicopter* only showed superficial connection with the topics in question and it was thus concluded that little or no aircraft weapons work is being done at NRC. As the F-18 is Canada's primary fighter aircraft, the term was used *as* a probe to confirm the result of the search. Nothing was returned, confirming the earlier inference.

CONCLUSIONS

The preliminary tests seem to indicate that analysts might find VITA useful to research major corpora and to confirm or dis-confirm information concerning the activities described in that corpus and, second, their conceptual inter-relations.

Further experimentation is required to determine the applicability of the VITA as a tool for Tech Watch. This could be accomplished by a targeted search of a structured database using sections of a more detailed taxonomy or thesaurus.

The delta version is now installed in small practical applications, for user testing. Following several months, the development team will select features for inclusion in the C++ [epsilon] version, for production use.

SYMPOSIA DISCUSSION – PAPER NO: 16

Author's Name: Dr. Zack Jacobson, Health Canada, Canada

Question:

In the search process does the user apply ontology?

Author's Response:

No, there are some built in, but the real ontology is the clustering.

Question:

Does the application use one search engine at once?

Author's Response:

In theory the number of search engines used is limited only by processing power.

Question:

Do you see a value in using multiple engines?

Author's Response:

A large difference has not been noticed. Google has been found to produce good results. Hummingbird has also been used with this system, and is a good alternative as it also returns good results and the html is stable.

Comment:

It is interesting to see the 3D connections, which have many similarities to the latest brain research. This system could have many applications where the associations and connections within a dataset is the important information to be brought out.



Text Mapping for Technology Watch

A research application

Z. Jacobson, Susan McIntyre, Tiit Romet, CA

Outline

- ◆ Background,
- ◆ Why Tech Watch,
- ◆ VITA,
- ◆ the tests,
- ◆ results,
- ◆ what next.

Can we engineer?

- ◆ Everyone's expectations from visual search are changing rapidly.
- ◆ Everyone brings a different mindset to the table.
- ◆ Take a configuration as starting point; use it!



Why Tech Watch?

Consider technologies by impact.

- Automobiles
- Cell phones
- Internet, Interstates
- P.C.
- ◆ The greatest effects are not easily predicted!
 - ◆ How to do it?

An answer, possibly

- ◆ Use large, parallel input mode
--vision.

i.e., Convert problem to navigation among
elements and features in space

e.g., driving in traffic; swimming underwater

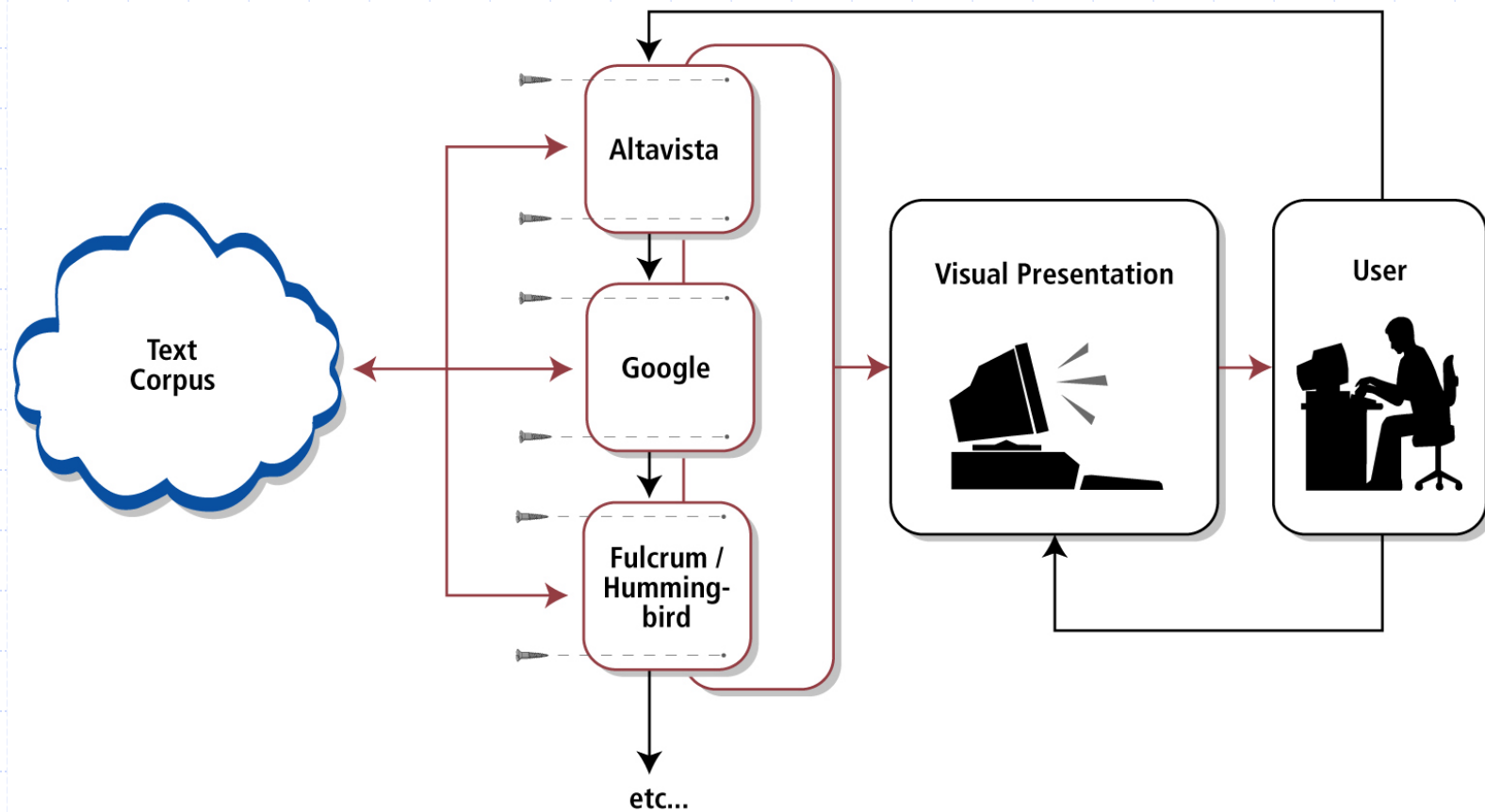
VITA - a visual front end for document search/management systems

Research testbed

- ◆ Search under user control
- ◆ Results-presentation under user control
- ◆ Search engine independent
- ◆ Various prototypes

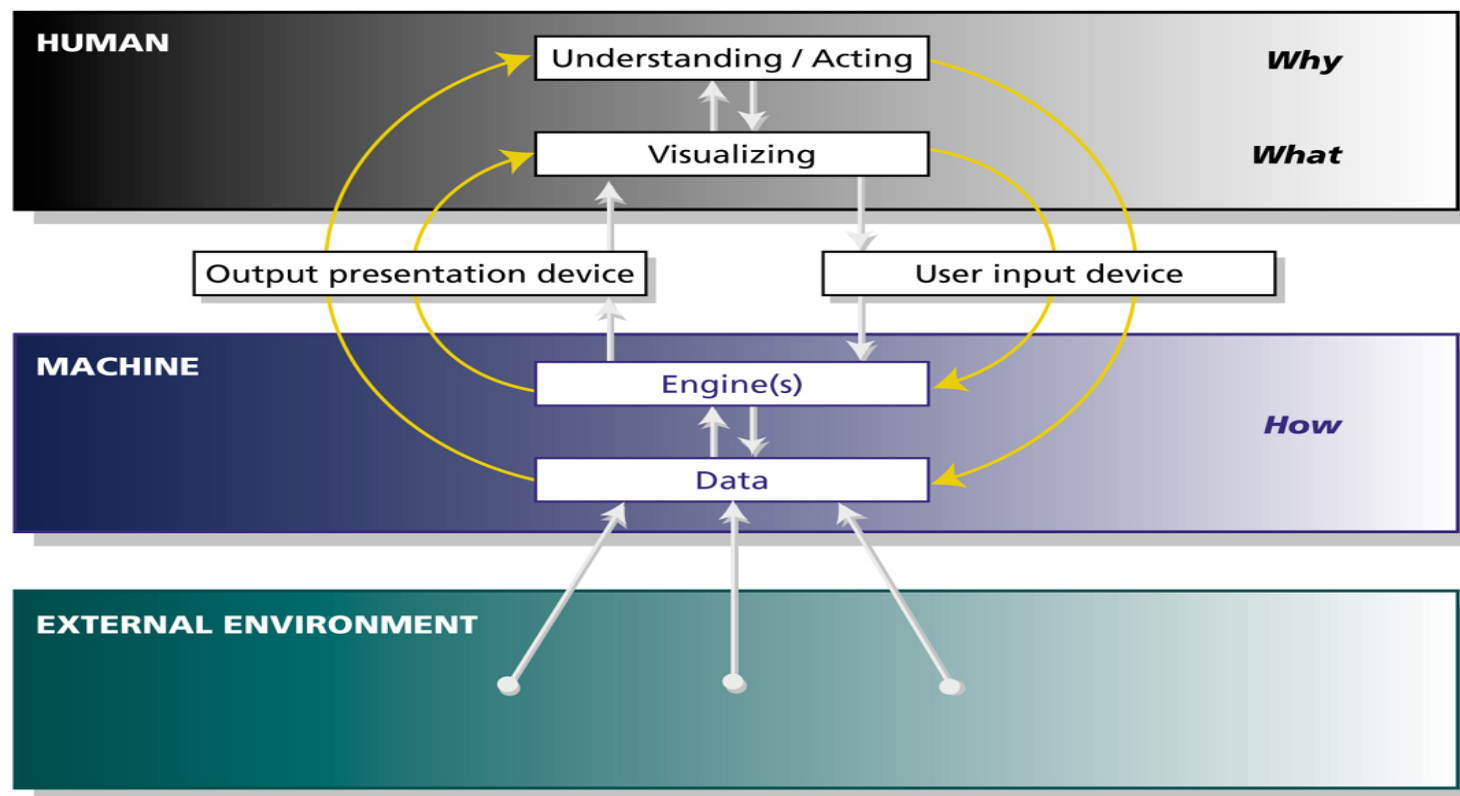
Standard interface awaiting completion to allow complete separation from the underlying search mechanism.

VITA concept



"Reference model"

Interpreting Visualization in Massive Datasets



Fielded instantiations

- ◆ Health Canada
 - ◆ WHO health watch INTEL and early warning online service
- ◆ [ex] DERA Malvern
 - γ version
- ◆ CA IO lab
 - ◆ Test on clustering hacked DIN's
- ◆ Other, various
 - ◆ Zack, Randy for websearch
 - ◆ Under formal evaluation

Two ways to view VITA

- ◆ Tool to discover the conceptual relations among elements of text in a massive corpus.
- OR
- ◆ Tool to help in reducing document search complexity

Technology watch work largely the former.

VITA Prototypes used for TW

◆ VITA- Δ

- Written in Visual Basic 6.0
- Based on capabilities of VITA- β & γ
- Research testbed for visual and control side
 - ◆ Parametrically configurable at run time
 - ◆ Default settings
- Interfaced to multiple search engines
 - ◆ Approx. 2 days effort to interface a new search engine
- TW use—topics a few at a time.

VITA Prototypes used for TW

◆ VITA-ε

- Written in C++
- Based on capabilities of VITA-β, as amended.
- Faster, more robust than VITA-Δ; no 3rd party dependencies.
- Modular construction
 - ◆ Incomplete but still usable with limited clustering and handling
- TW uses—many-to-many mappings of topics with documents/activities.

Elements that affect results

- ◆ Watcher's style and topic chosen
- ◆ Search domain
- ◆ Search engine
- ◆ Search technique

Intersecting elements

- ◆ TW topic chosen

Broad or narrow

- ◆ Appropriate search domain

Technology taxonomy

- ◆ Search engine

Choice of several

- ◆ Search technique

Naïve or sophisticated

Examples

- ◆ Topic for Technology Watch

from the DTAP [aircraft weapons, engine, e.g.]

- ◆ Appropriate search domain

NRC [or NCE, CIA, Jane's,]

- ◆ Search engine

Google [or Alta Vista, Fulcrum,]

- ◆ Search technique

Naïve or sophisticated [e.g., familiar with engine?]

Two examples to check an NRC...

Air Platforms (DTAP BT)*

- UF* Air Weapons Systems (CA Thrust 13e)
 - Aircraft/Weapon System Compatibility (CA Project 13ec)

Fixed-Wing Vehicles (DTAP NT)*

Rotary-Wing Vehicles (DTAP NT)

Integrated High-Performance Turbine Engine Technologies (DTAP NT)

Aircraft Power (DTAP NT)

- UF - Advanced Power Sources (CA Project 13gf)
 - Advanced Portable Fuel Cells (CA Project 13gj)

High-Speed Propulsion and Fuels (DTAP NT)

...

On the NRC site, with VITA- Δ , using Google. Example 1

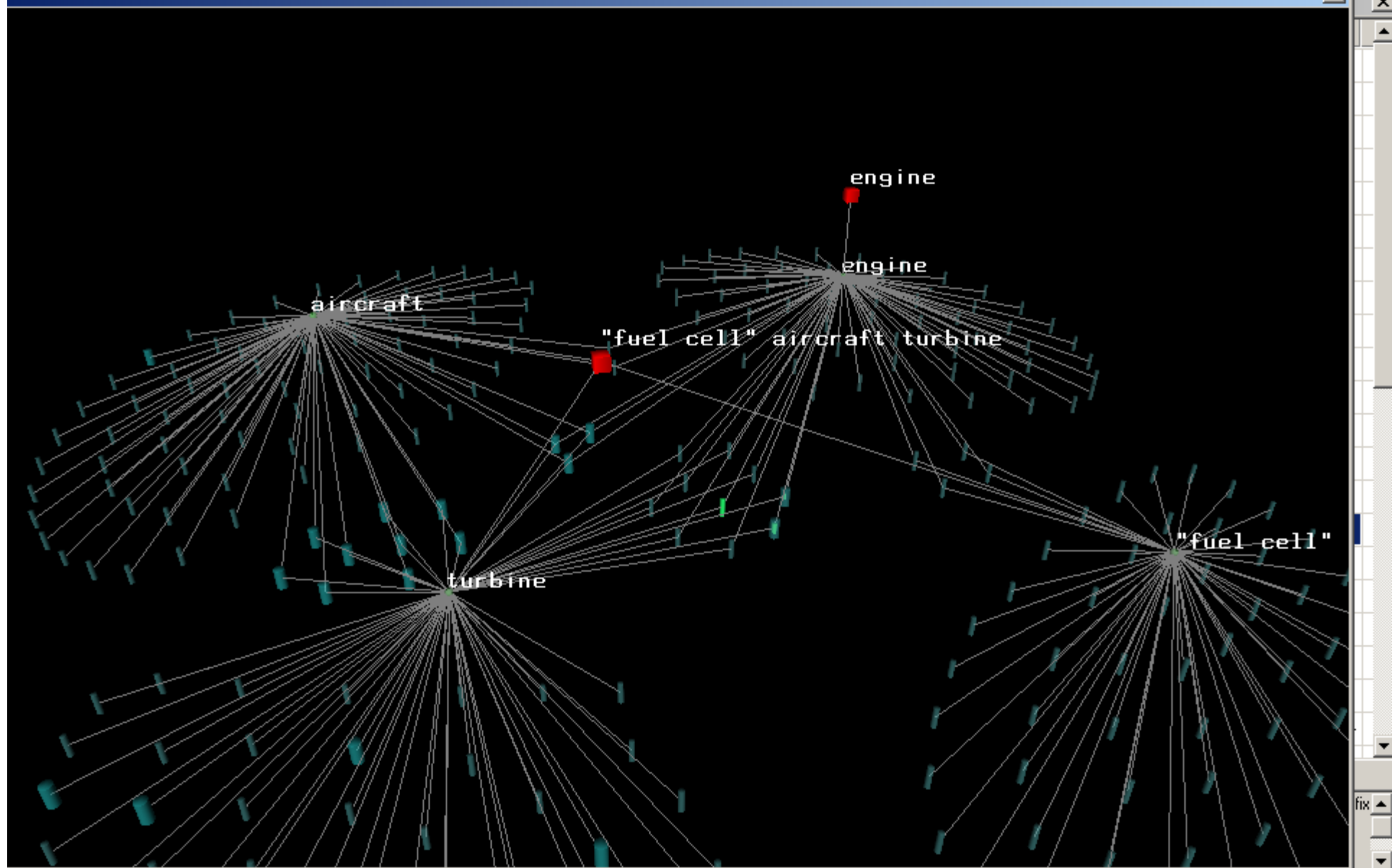
- ◆ Sequence of queries
 - Aircraft, “fixed-wing”, helicopter
 - ◆ Yields a filled field
 - Weapon weapons system
 - ◆ Shows little intersection, none with fixed wing
 - ◆ Scan the intersects with helicopters
- ◆ Conclude—little or no aircraft weapons work at NRC.
 - F-18
 - ◆ Probe query to confirm.

Search Terms: engine

Add Term(s)

Add Query

View Results



On the NRC site, with VITA, using Google. Example 2

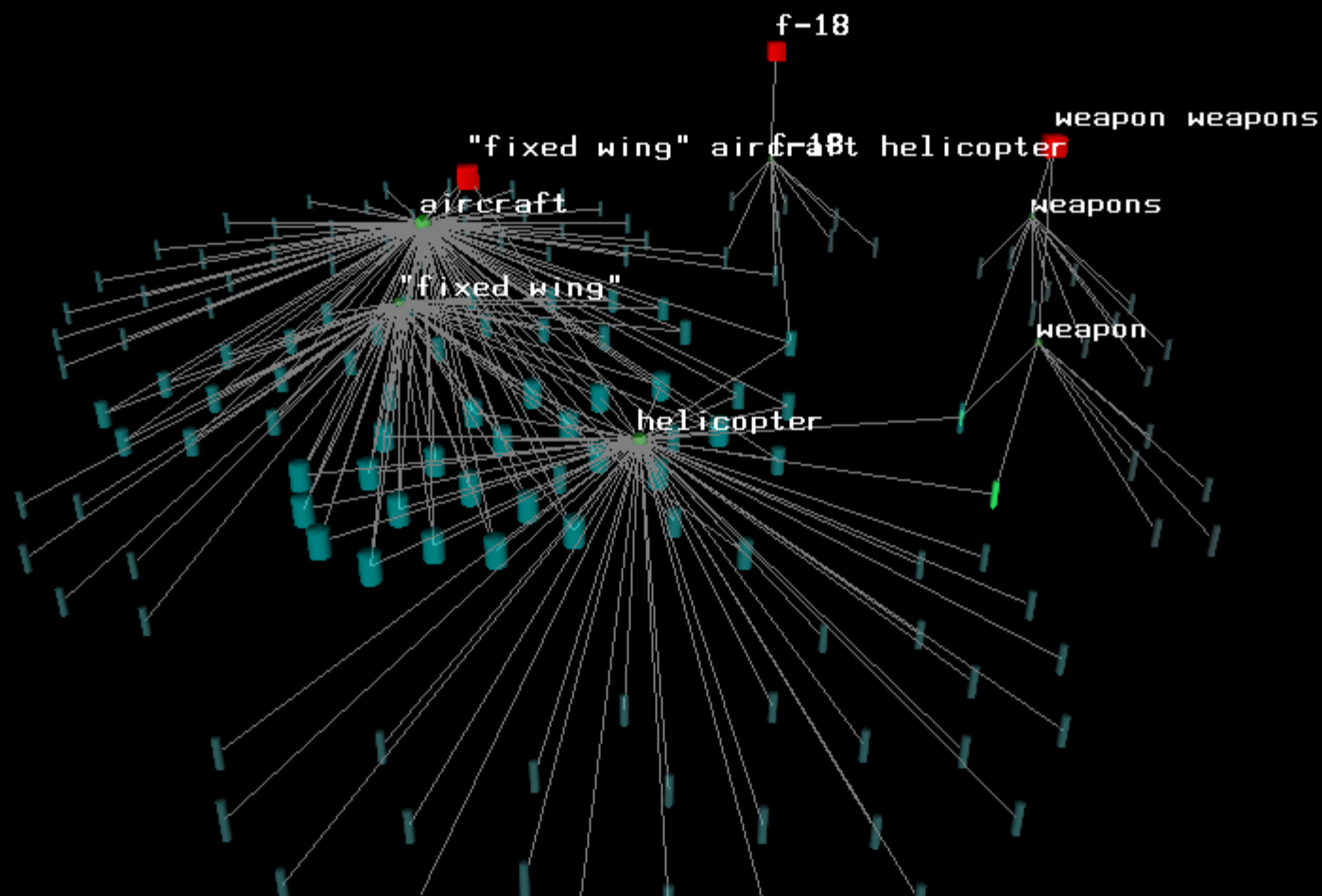
- ◆ Sequence of queries
 - Aircraft, turbine engines
 - ◆ Yields a filled field, engines unnecessary--removed
 - Fuel cell
 - ◆ We see elements, but no intersection to aircraft
- ◆ Conclude—no aircraft fuel cell work at NRC
 - engine
 - ◆ Probe query, reinserted to confirm.
- ◆ Disconfirmed? Scan intersect “hits”!
 - Elements found for aircraft engines, for aircraft and engines, but none for engines as powered by fuel cells.

Search Terms: f-18

Add Term(s)

Add Query

View Results



VITA-ε map of the DTAP

- Defence taxonomy elements
 - ◆ Term layer

Against

- Canadian Defence activities
 - ◆ Hit layer

